An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Make a Big Difference?

Ryan Giordano
MIT

Rachael Meager
LSE

Tamara Broderick
MIT

Job talk 2021

Dropping data: Motivation

You're a data analyst, and you've

- Gathered some exchangeable data,
- Cleaned up / removed outliers,
- Checked for correct specification, and
- Drawn a conclusion from your statistical analysis
  (e.g., based the sign / significance of some estimated parameter).

Dropping data: Motivation

You're a data analyst, and you've

- Gathered some exchangeable data,
- Cleaned up / removed outliers,
- Checked for correct specification, and
- Drawn a conclusion from your statistical analysis (e.g., based the sign / significance of some estimated parameter).

Well done!

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change both	15	7.03 (2.55)

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change both	15	7.03 (2.55)

By removing very few data points ($15/16560 \approx 0.1\%$), we can reverse the qualitative conclusions of the original study!

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change both	15	7.03 (2.55)

By removing very few data points ($15/16560 \approx 0.1\%$), we can reverse the qualitative conclusions of the original study!
Question: Is the reported interval $-4.55 \pm (5.88)$ a reasonable description of the uncertainty in the estimated efficacy of microcredit?

Dropping data: Motivation

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Dropping data: Motivation

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?
Not always!

Dropping data: Motivation

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Not always!

...but sometimes, surely yes.

For example, often in economics:

- Small fractions of data are missing not-at-random,
- Policy population is different from analyzed population,
- We report a convenient summary (e.g. mean) of a complex effect,
- Models are stylized proxies of reality.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?
The number of subsets $\binom{N}{\lfloor \alpha N \rfloor}$ can be very large even when α is very small.
In the MX microcredit study, $\binom{16560}{15} \approx 1.4 \cdot 10^{51}$ sets to check for $\alpha = 0.0009$.
We provide a fast, automatic approximation based on the **influence function**.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?
Non-robustness to removal of $\lfloor \alpha N \rfloor$ points is:

- Not (necessarily) caused by misspecification.
- Not (necessarily) caused by outliers.
- Not captured by standard errors.
- Not mitigated by large N.
- Primarily determined by the **signal to noise** ratio

  ... in a sense which we will define.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Question 3: When is our approximation accurate?

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Question 3: When is our approximation accurate?

- We provide deterministic error bounds for small α.
- We show the accuracy in simple experiments.
- We show the accuracy in a number of real-world experiments.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Question 3: When is our approximation accurate?

Conclusion: Related work and future directions

Question 1:
How do we find influential datapoints?

Which estimators do we study?

Suppose we have N data points $d_1, \ldots, d_N$. Then:

$$\hat{\theta} := \vec{\theta} \text{ such that } \sum_{n=1}^{N} G(\vec{\theta}, d_n) = 0_P.$$

Leave points out by setting their elements of $\vec{w}$ to zero.
These are "Z-estimators," i.e., roots of estimating equations.
Examples: all minimizers of empirical loss (OLS, MLE, VB), and more.

Which estimators do we study?

Suppose we have N data points $d_1, \ldots, d_N$. Then:

$$\hat{\theta}(\vec{w}) := \vec{\theta} \text{ such that } \sum_{n=1}^{N} \vec{w}_n G(\vec{\theta}, d_n) = 0_P.$$

Leave points out by setting their elements of $\vec{w}$ to zero.
These are "Z-estimators," i.e., roots of estimating equations.
Examples: all minimizers of empirical loss (OLS, MLE, VB), and more.

Which estimators do we study?

Suppose we have N data points $d_1, \ldots, d_N$. Then:

$$\hat{\theta}(\vec{w}) := \vec{\theta} \text{ such that } \sum_{n=1}^{N} \vec{w}_n G(\vec{\theta}, d_n) = 0_P.$$

Leave points out by setting their elements of $\vec{w}$ to zero.

Fix a quantity of interest, $\phi(\vec{\theta})$.
Let the **"signal"**, Δ, be a "large" change in ϕ.

Which estimators do we study?

Suppose we have N data points $d_1, \ldots, d_N$. Then:

$$\hat{\theta}(\vec{w}) := \vec{\theta} \text{ such that } \sum_{n=1}^{N} \vec{w}_n G(\vec{\theta}, d_n) = 0_P.$$

Leave points out by setting their elements of $\vec{w}$ to zero.

Fix a quantity of interest, $\phi(\vec{\theta})$.
Let the **"signal"**, Δ, be a "large" change in ϕ.

Examples:

$$\phi(\vec{\theta}) = \vec{\theta}_p$$

$$\phi(\vec{\theta}) = \vec{\theta}_p + \frac{1.96}{\sqrt{N}} \hat{\sigma}_\phi(\vec{\theta})$$

Which estimators do we study?

Suppose we have N data points $d_1, \ldots, d_N$. Then:

$$\hat{\theta}(\vec{w}) := \vec{\theta} \text{ such that } \sum_{n=1}^{N} \vec{w}_n G(\vec{\theta}, d_n) = 0_P.$$

Leave points out by setting their elements of $\vec{w}$ to zero.

Fix a quantity of interest, $\phi(\vec{\theta})$.
Let the **"signal"**, Δ, be a "large" change in ϕ.

Examples:

$$\phi(\vec{\theta}) = \vec{\theta}_p$$

$$\phi(\vec{\theta}) = \vec{\theta}_p + \frac{1.96}{\sqrt{N}} \hat{\sigma}_\phi(\vec{\theta})$$

Which estimators do we study?

Suppose we have N data points $d_1, \ldots, d_N$. Then:

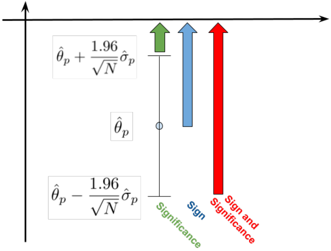
$$\hat{\theta}(\vec{w}) := \vec{\theta} \ \text{ such that } \ \sum_{n=1}^{N} \vec{w}_n G(\vec{\theta}, d_n) = 0_P.$$

Leave points out by setting their elements of $\vec{w}$ to zero.

Fix a quantity of interest, $\phi(\vec{\theta})$.
Let the **"signal"**, Δ, be a "large" change in ϕ.

Examples:

$$\phi(\vec{\theta}) = \vec{\theta}_p$$
$$\phi(\vec{\theta}) = \vec{\theta}_p + \frac{1.96}{\sqrt{N}} \hat{\sigma}_\phi(\vec{\theta})$$



Which estimators do we study?

Suppose we have N data points $d_1, \ldots, d_N$. Then:

$$\hat{\theta}(\vec{w}) := \vec{\theta} \ \text{ such that } \ \sum_{n=1}^{N} \vec{w}_n G(\vec{\theta}, d_n) = 0_P.$$

Leave points out by setting their elements of $\vec{w}$ to zero.

Fix a quantity of interest, $\phi(\vec{\theta})$.
Let the **"signal"**, Δ, be a "large" change in ϕ.

Can we reverse our conclusion by dropping $\lfloor \alpha N \rfloor$ datapoints?

Which estimators do we study?

Suppose we have N data points $d_1, \ldots, d_N$. Then:

$$\hat{\theta}(\vec{w}) := \vec{\theta} \text{ such that } \sum_{n=1}^{N} \vec{w}_n G(\vec{\theta}, d_n) = 0_P.$$

Leave points out by setting their elements of $\vec{w}$ to zero.

Fix a quantity of interest, $\phi(\vec{\theta})$.
Let the **"signal"**, Δ, be a "large" change in ϕ.

Can we reverse our conclusion by dropping $\lfloor \alpha N \rfloor$ datapoints?

$\Leftrightarrow$

Is there a $\vec{w}$, with $\lfloor \alpha N \rfloor$ zeros, such that $\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \geq \Delta$?

Hard! Evaluating $\hat{\theta}(\vec{w})$ is costly and lots of $\vec{w}$ have $\lfloor \alpha N \rfloor$ zeros.

Taylor series approximation.

Is there a $\vec{w}$, with $\lfloor \alpha N \rfloor$ zeros, such that $\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \geq \Delta$?

To simplify the search over $\vec{w}$, we form the Taylor series approximation:

$$\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \approx \phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}) := -\sum_{n:\vec{w}_n=0} \psi_n, \text{ where } \psi_n := \left.\frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n}\right|_{\vec{1}}.$$

Taylor series approximation.

Is there a $\vec{w}$, with $\lfloor \alpha N \rfloor$ zeros, such that $\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \geq \Delta$?

To simplify the search over $\vec{w}$, we form the Taylor series approximation:

$$\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \approx \phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}) := -\sum_{n:\vec{w}_n=0} \psi_n, \text{ where } \psi_n := \left.\frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n}\right|_{\vec{1}}.$$

The values ψ_n are the **"empirical influence function."** [Hampel, 1986]
The ψ_n can be **easily and automatically** computed from $\hat{\theta}$.
The approximation is **typically accurate** for small α.

Taylor series approximation.

Is there a $\vec{w}$, with $\lfloor \alpha N \rfloor$ zeros, such that $\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \geq \Delta$?

To simplify the search over $\vec{w}$, we form the Taylor series approximation:

$$\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \approx \phi^{\text{lin}}(\vec{w}) - \phi(\hat{\theta}) := -\sum_{n:\vec{w}_n=0} \psi_n, \text{ where } \psi_n := \left.\frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n}\right|_{\vec{1}}.$$

The values ψ_n are the **"empirical influence function."** [Hampel, 1986]
The ψ_n can be **easily and automatically** computed from $\hat{\theta}$.
The approximation is **typically accurate** for small α.

Is there a $\vec{w}$, with $\lfloor \alpha N \rfloor$ zeros, such that $\phi^{\text{lin}}(\vec{w}) - \phi(\hat{\theta}) \geq \Delta$?

Taylor series approximation.

Is there a $\vec{w}$, with $\lfloor \alpha N \rfloor$ zeros, such that $\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \geq \Delta$?

To simplify the search over $\vec{w}$, we form the Taylor series approximation:

$$\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \approx \phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}) := - \sum_{n : \vec{w}_n = 0} \psi_n, \text{ where } \psi_n := \left. \frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n} \right|_{\vec{1}}.$$

The values ψ_n are the **"empirical influence function."** [Hampel, 1986]
The ψ_n can be **easily and automatically** computed from $\hat{\theta}$.
The approximation is **typically accurate** for small α.

Is there a $\vec{w}$, with $\lfloor \alpha N \rfloor$ zeros, such that $\phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}) \geq \Delta$?

Easy! The most influential points for $\phi^{\mathrm{lin}}(\vec{w})$ have the most negative ψ_n.

Computing the influence function.

How to compute $\psi_n := \left.\frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n}\right|_{\vec{1}}$? Recall $\sum_{n=1}^{N} \vec{w}_n G(\hat{\theta}(\vec{w}), d_n) = 0_P$.

Step zero: Implement software to compute $G(\theta, d_n)$ and $\phi(\theta)$. Find $\hat{\theta}$.

Step one: By the chain rule, $\psi_n = \left.\frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n}\right|_{\vec{1}} = \left.\frac{\mathrm{d}\phi(\theta)}{\mathrm{d}\theta^T}\right|_{\hat{\theta}} \left.\frac{\partial \hat{\theta}(\vec{w})}{\partial \vec{w}_n}\right|_{\vec{1}}$.

Step two: By the implicit function theorem:

$$\left.\frac{\partial \hat{\theta}(\vec{w})}{\partial \vec{w}_n}\right|_{\vec{1}} = \frac{1}{N}\left(\frac{1}{N}\sum_{n'=1}^{N} \frac{\partial}{\partial \theta^T} G(\vec{\theta}, d_{n'})\Big|_{\hat{\theta}}\right)^{-1} G(\hat{\theta}, d_n).$$

Step three: Use *automatic differentiation* on $\phi(\theta)$ and $G(\theta, d_n)$ from step zero to compute $\frac{\partial \phi(\theta)}{\partial \theta^T}$ and $\frac{\partial}{\partial \theta^T} G(\vec{\theta}, d_n)$.

- The user does step zero. The rest is automatic.
- The primary computational expense is the Hessian inverse.
- Automatic differentiation is the chain rule applied to a program.
- Typically $\psi_n = O(N^{-1})$.

Taylor series approximation.

Procedure:

Taylor series approximation.

Procedure:

1. Compute the "original" estimator, $\hat{\theta}$ and $\phi(\hat{\theta})$.

Taylor series approximation.

Procedure:

1. Compute the “original” estimator, $\hat{\theta}$ and $\phi(\hat{\theta})$.
2. Compute and sort the influence scores, $\psi_{(1)} \leq \psi_{(2)} \leq \ldots \leq \psi_{(N)}$.

Taylor series approximation.

Procedure:

1. Compute the "original" estimator, $\hat{\theta}$ and $\phi(\hat{\theta})$.
2. Compute and sort the influence scores, $\psi_{(1)} \leq \psi_{(2)} \leq \ldots \leq \psi_{(N)}$.
3. Let $\vec{w}^*$ leave out the data corresponding to $\psi_{(1)}, \ldots, \psi_{(\lfloor \alpha N \rfloor)}$.

Taylor series approximation.

Procedure:

1. Compute the "original" estimator, $\hat{\theta}$ and $\phi(\hat{\theta})$.
2. Compute and sort the influence scores, $\psi_{(1)} \leq \psi_{(2)} \leq \ldots \leq \psi_{(N)}$.
3. Let $\vec{w}^*$ leave out the data corresponding to $\psi_{(1)}, \ldots, \psi_{(\lfloor \alpha N \rfloor)}$.
4. Report non-robustness if $\Delta \leq \phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = -\sum_{n=1}^{\lfloor \alpha N \rfloor} \psi_{(n)}$.

Taylor series approximation.

Procedure:

1. Compute the "original" estimator, $\hat{\theta}$ and $\phi(\hat{\theta})$.
2. Compute and sort the influence scores, $\psi_{(1)} \leq \psi_{(2)} \leq \ldots \leq \psi_{(N)}$.
3. Let $\vec{w}^*$ leave out the data corresponding to $\psi_{(1)}, \ldots, \psi_{(\lfloor \alpha N \rfloor)}$.
4. Report non-robustness if $\Delta \leq \phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = -\sum_{n=1}^{\lfloor \alpha N \rfloor} \psi_{(n)}$.
5. **Optional:** Compute $\hat{\theta}(\vec{w}^*)$, and verify that $\Delta \leq \phi(\hat{\theta}(\vec{w}^*)) - \phi(\hat{\theta})$.

Question 2:
What makes an estimator non-robust?

What makes an estimator non-robust?

For $N = 5,000$ data points, compute the OLS estimator from:

Regressors	Residuals	Responses
$x_n \sim \mathcal{N}(0, \sigma_x^2)$	$\varepsilon_n \sim \mathcal{N}(0, \sigma_\varepsilon^2)$	$y_n = \theta_0 x_n + \varepsilon_n$

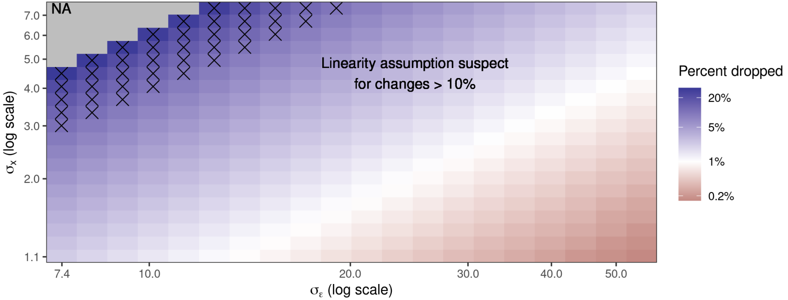


Figure: The approximate perturbation inducing proportion at differing values of σ_x and σ_ε. Red colors indicate datasets whose sign can is predicted to change when dropping less than 1% of datapoints. The grey areas indicate $\hat{\Psi}_\alpha = \mathtt{NA}$, a failure of the linear approximation to locate any way to change the sign.

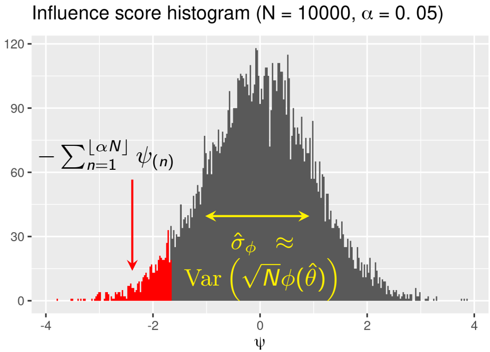
What makes an estimator non-robust? A tail sum.

Report non-robustness if:

$$\Delta \leq \phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = -\sum_{n=1}^{\lfloor \alpha N \rfloor} \psi_{(n)} =: \hat{\sigma}_\phi \hat{\mathscr{T}}_\alpha$$

We will show that:

- The "noise" $\hat{\sigma}_\phi^2 \to \mathrm{Var}(\sqrt{N}\phi)$ [Hampel, 1986]
- The "shape" $\hat{\mathscr{T}}_\alpha \leq \sqrt{\alpha(1-\alpha)}$ and converges to a nonzero constant



Three steps:

1. $\tilde{\psi}_n := N\psi_n$ has a non-degenerate distribution.
2. $\hat{\sigma}_\phi := \frac{1}{N}\sum_{n=1}^{N} \tilde{\psi}_n^2$ estimates $\mathrm{Var}\left(\sqrt{N}\phi(\hat{\theta})\right)$.
3. $\hat{\mathscr{T}}_\alpha := \frac{-\frac{1}{N}\sum_{n=1}^{\lfloor \alpha N \rfloor} \tilde{\psi}_{(n)}}{\hat{\sigma}_\phi} \leq \sqrt{\alpha(1-\alpha)}$ and converges to a constant $\neq 0$.

Three steps:

1. $\tilde{\psi}_n := N\psi_n$ has a non-degenerate distribution.

Assume that $\hat{\theta} \xrightarrow{p} \theta_\infty$ and laws of large numbers apply.
By direct computation,

$$\tilde{\psi}_n = N\psi_n = \underbrace{\left.\frac{\mathrm{d}\phi(\theta)}{\mathrm{d}\theta^T}\right|_{\hat{\theta}}}_{\xrightarrow{p} \left.\frac{\mathrm{d}\phi(\theta)}{\mathrm{d}\theta^T}\right|_{\theta_\infty}} \underbrace{\left(\frac{1}{N}\sum_{n'=1}^{N} \frac{\partial}{\partial\theta^T} G(\vec{\theta}, d_{n'})\Big|_{\hat{\theta}}\right)^{-1}}_{\xrightarrow{p} \mathbb{E}_d\left[\frac{\partial}{\partial\theta^T} G(\vec{\theta}, d)\Big|_{\theta_\infty}\right]} \underbrace{G(\hat{\theta}, d_n)}_{\xrightarrow{p} G(\theta_\infty, d_n)}.$$

It follows that $\tilde{\psi}_n$ have a non-degenerate distribution for all N.

Three steps:

1. $\tilde{\psi}_n := N\psi_n$ has a non-degenerate distribution.
2. $\hat{\sigma}_\phi := \frac{1}{N}\sum_{n=1}^{N}\tilde{\psi}_n^2$ estimates $\mathrm{Var}\left(\sqrt{N}\phi(\hat{\theta})\right)$.

Argument 1: A linear approximation to the bootstrap.
Let $\mathrm{Boot}(\vec{w})$ denote the distribution of random bootstrap weights.

$$
\begin{aligned}
\underset{\mathrm{Boot}(\vec{w})}{\mathrm{Var}}\left(\sqrt{N}\phi(\hat{\theta})\right) &\approx \underset{\mathrm{Boot}(\vec{w})}{\mathrm{Var}}\left(\sqrt{N}\phi^{\mathrm{lin}}(\hat{\theta})\right) \\
&= \underset{\mathrm{Boot}(\vec{w})}{\mathrm{Var}}\left(\sqrt{N}\sum_{n=1}^{N}\psi_n(\vec{w}_n - 1)\right) \\
&= \sum_{n=1}^{N} N\psi_n^2 = \frac{1}{N}\sum_{n=1}^{N}\tilde{\psi}_n^2 = \hat{\sigma}_\phi^2.
\end{aligned}
$$

Argument 2: Formally, $\hat{\sigma}_\phi^2$ is the "sandwich covariance" estimator [Huber, 1967, Stefanski and Boos, 2002].
Argument 3: Influence functions and von Mises calculus [Mises, 1947, Reeds, 1976].

Three steps:

1. $\tilde{\psi}_n := N\psi_n$ has a non-degenerate distribution.
2. $\hat{\sigma}_\phi := \frac{1}{N}\sum_{n=1}^{N} \tilde{\psi}_n^2$ estimates $\mathrm{Var}\left(\sqrt{N}\phi(\hat{\theta})\right)$.
3. $\hat{\mathscr{T}}_\alpha := \frac{-\frac{1}{N}\sum_{n=1}^{\lfloor \alpha N \rfloor} \tilde{\psi}_{(n)}}{\hat{\sigma}_\phi} \le \sqrt{\alpha(1-\alpha)}$ and converges to a constant $\neq 0$.

By definition,

$$-\sum_{n=1}^{\lfloor \alpha N \rfloor} \psi_{(n)} =: \hat{\sigma}_\phi \hat{\mathscr{T}}_\alpha \qquad \Rightarrow \qquad \hat{\mathscr{T}}_\alpha = -\frac{1}{N}\sum_{n=1}^{\lfloor \alpha N \rfloor} \frac{\tilde{\psi}_{(n)}}{\hat{\sigma}_\phi}.$$

By Cauchy-Schwartz,

$$\hat{\mathscr{T}}_\alpha \le \underbrace{\left(\frac{1}{N}\sum_{n=1}^{N} \frac{\tilde{\psi}_n^2}{\hat{\sigma}_\phi^2}\right)^{1/2}}_{=1} \left(\frac{1}{N}\sum_{n=1}^{N} \mathbb{I}\left(n \le \alpha N\right)^2\right)^{1/2} \le \sqrt{\alpha}$$

A slightly more careful analysis which accounts for the fact that $\sum_{n=1}^{N} \psi_n = 0$ gives $\hat{\mathscr{T}}_\alpha \le \sqrt{\alpha(1-\alpha)}$.

Corollaries.

Report non-robustness if:

$$\Delta \leq \phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = \hat{\sigma}_\phi \hat{\mathscr{T}}_\alpha \qquad \Leftrightarrow \qquad \frac{\Delta}{\hat{\sigma}_\phi} \leq \hat{\mathscr{T}}_\alpha.$$

We call $\frac{\Delta}{\hat{\sigma}_\phi}$ the "signal to noise ratio."

Corollaries.

Report non-robustness if:

$$\Delta \leq \phi^{\text{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = \hat{\sigma}_\phi \hat{\mathscr{T}}_\alpha \qquad \Leftrightarrow \qquad \frac{\Delta}{\hat{\sigma}_\phi} \leq \hat{\mathscr{T}}_\alpha.$$

We call $\frac{\Delta}{\hat{\sigma}_\phi}$ the “signal to noise ratio.”

Corollary: Non-robustness possible even with correct specification.

Corollaries.

Report non-robustness if:

$$\Delta \leq \phi^{\text{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = \hat{\sigma}_\phi \hat{\mathscr{T}}_\alpha \qquad \Leftrightarrow \qquad \frac{\Delta}{\hat{\sigma}_\phi} \leq \hat{\mathscr{T}}_\alpha.$$

We call $\frac{\Delta}{\hat{\sigma}_\phi}$ the "signal to noise ratio."

Corollary: Non-robustness possible even with correct specification.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out robustness does not vanish as $N \to \infty$.

Corollaries.

Report non-robustness if:

$$\Delta \le \phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = \hat{\sigma}_\phi \hat{\mathscr{T}}_\alpha \qquad \Leftrightarrow \qquad \frac{\Delta}{\hat{\sigma}_\phi} \le \hat{\mathscr{T}}_\alpha.$$

We call $\frac{\Delta}{\hat{\sigma}_\phi}$ the "signal to noise ratio."

Corollary: Non-robustness possible even with correct specification.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out robustness does not vanish as $N \to \infty$.

Recall that standard errors reject when $\frac{\Delta}{\hat{\sigma}_\phi} \le \frac{1.96}{\sqrt{N}}$.

Corollaries.

Report non-robustness if:

$$\Delta \leq \phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = \hat{\sigma}_\phi \hat{\mathscr{T}}_\alpha \qquad \Leftrightarrow \qquad \frac{\Delta}{\hat{\sigma}_\phi} \leq \hat{\mathscr{T}}_\alpha.$$

We call $\frac{\Delta}{\hat{\sigma}_\phi}$ the "signal to noise ratio."

Corollary: Non-robustness possible even with correct specification.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out robustness does not vanish as $N \to \infty$.

Recall that standard errors reject when $\frac{\Delta}{\hat{\sigma}_\phi} \leq \frac{1.96}{\sqrt{N}}$.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out is different from standard errors.

Corollaries.

Report non-robustness if:

$$\Delta \leq \phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = \hat{\sigma}_\phi \hat{\mathscr{T}}_\alpha \qquad \Leftrightarrow \qquad \frac{\Delta}{\hat{\sigma}_\phi} \leq \hat{\mathscr{T}}_\alpha.$$

We call $\frac{\Delta}{\hat{\sigma}_\phi}$ the "signal to noise ratio."

Corollary: Non-robustness possible even with correct specification.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out robustness does not vanish as $N \to \infty$.

Recall that standard errors reject when $\frac{\Delta}{\hat{\sigma}_\phi} \leq \frac{1.96}{\sqrt{N}}$.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out is different from standard errors.

Corollary: Insignificance is always non-robust.
Take $\Delta = \frac{1.96 \hat{\sigma}_\phi}{\sqrt{N}} \to 0 \leq \hat{\mathscr{T}}_\alpha$.

Corollaries.

Report non-robustness if:

$$\Delta \le \phi^{\text{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = \hat{\sigma}_\phi \hat{\mathscr{T}}_\alpha \qquad \Leftrightarrow \qquad \frac{\Delta}{\hat{\sigma}_\phi} \le \hat{\mathscr{T}}_\alpha.$$

We call $\frac{\Delta}{\hat{\sigma}_\phi}$ the "signal to noise ratio."

Corollary: Non-robustness possible even with correct specification.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out robustness does not vanish as $N \to \infty$.

Recall that standard errors reject when $\frac{\Delta}{\hat{\sigma}_\phi} \le \frac{1.96}{\sqrt{N}}$.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out is different from standard errors.

Corollary: Insignificance is always non-robust.
Take $\Delta = \frac{1.96 \hat{\sigma}_\phi}{\sqrt{N}} \to 0 \le \hat{\mathscr{T}}_\alpha$.

Corollary: Gross outliers primarily affect robustness through $\hat{\sigma}_\phi$.
Cauchy-Schwartz is tight when all the influence scores are the same.

Question 3:
When is our approximation accurate?

The linear approximation.

For $N = 5,000$ data points, compute the OLS estimator from:

Regressors	Residuals	Responses
$x_n \sim \mathcal{N}(0, \sigma_x^2)$	$\varepsilon_n \sim \mathcal{N}(0, \sigma_\varepsilon^2)$	$y_n = \theta_0 x_n + \varepsilon_n$

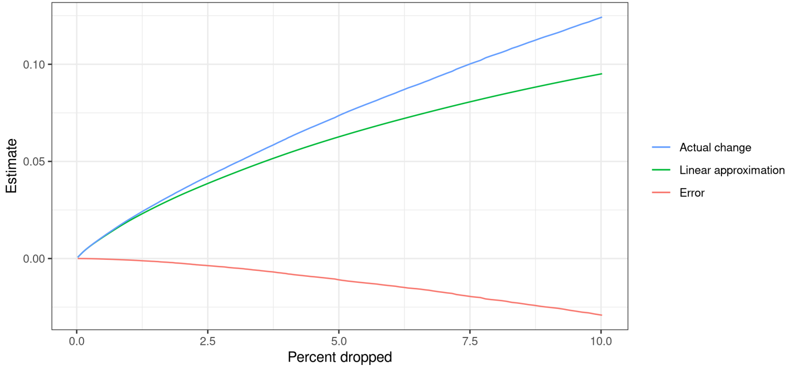
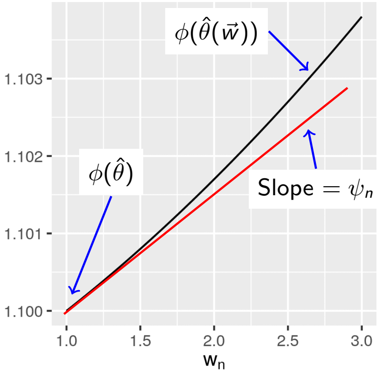


Figure: The actual change, linear approximation to the change, and approximation error.Here, $\sigma_x = 2$, $\sigma_\varepsilon = 1$, and $\theta_0 = 0.5$.

The linear approximation.



$$\phi(\hat{\theta}(\vec{w})) = \phi(\hat{\theta}) + \sum_{n=1}^{N} \psi_n(\vec{w}_n - 1) + \text{Higher-order derivatives}$$

Key idea: Controlling higher-order derivatives can control the error.

The linear approximation.

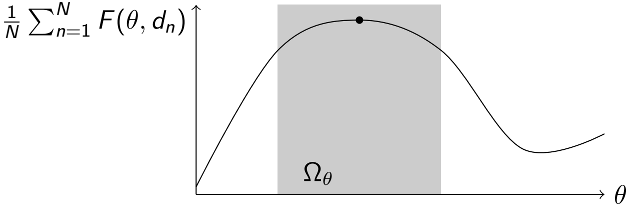
Let W_α be the set of weight vectors with no more than $\lfloor \alpha N \rfloor$ zeros.

Let $H(\theta, d_n) := \left. \frac{\partial G(\theta, d_n)}{\partial \theta^T} \right|_\theta$.

Assumption (Smooth Objective)

Fix the dataset. Assume there exists a compact $\Omega_\theta \subseteq \mathbb{R}^D$ with $\hat{\theta}(\vec{w}) \in \Omega_\theta$ for all $\vec{w} \in W_\alpha$. Assume that, for all $\theta \in \Omega_\theta$:

- $\frac{1}{N} \sum_{n=1}^N H(\theta, d_n)$ *and* $\frac{1}{N} \sum_{n=1}^N G(\theta, d_n)$ *are bounded.*
- $\frac{1}{N} \sum_{n=1}^N H(\theta, d_n)$ *is uniformly non-singular and Lipschitz (in* θ*).*
- $\phi(\theta)$ *has a Lipschitz first derivative.*



The linear approximation.

Theorem

Let Assumption 1 hold for a given dataset. Then there exists a sufficiently small α *such that*

$$\sup_{\vec{w}\in W_\alpha}\left|\phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}(\vec{w}))\right| \le C_1\alpha \;\; and \;\; \sup_{\vec{w}\in W_\alpha}\left|\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta})\right| \le C_2\sqrt{\alpha},$$

where C_1 *and* C_2 *are given by the quantities in the assumption.*

The linear approximation.

Theorem

Let Assumption 1 hold for a given dataset. Then there exists a sufficiently small α such that

$$\sup_{\vec{w} \in W_\alpha} \left| \phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}(\vec{w})) \right| \leq C_1 \alpha \ and \ \sup_{\vec{w} \in W_\alpha} \left| \phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \right| \leq C_2 \sqrt{\alpha},$$

where C_1 and C_2 are given by the quantities in the assumption.

Since $\alpha \ll \sqrt{\alpha}$ when α is small, Theorem 1 states that the linear approximation's error is of smaller order than the actual difference.

The linear approximation.

Theorem

Let Assumption 1 hold for a given dataset. Then there exists a sufficiently small α such that

$$\sup_{\vec{w} \in W_\alpha} \left| \phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}(\vec{w})) \right| \leq C_1 \alpha \ \textit{and} \ \sup_{\vec{w} \in W_\alpha} \left| \phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \right| \leq C_2 \sqrt{\alpha},$$

where C_1 and C_2 are given by the quantities in the assumption.

Proof sketch.

The second inequality follows from the smoothness of the objective. The first inequality follows from the smoothness of $d\hat{\theta}(\vec{w})/d\vec{w}$. □

The linear approximation.

Theorem

Let Assumption 1 hold for a given dataset. Then there exists a sufficiently small α *such that*

$$\sup_{\vec{w} \in W_\alpha} \left| \phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}(\vec{w})) \right| \le C_1 \alpha \;\; and \;\; \sup_{\vec{w} \in W_\alpha} \left| \phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \right| \le C_2 \sqrt{\alpha},$$

where C_1 *and* C_2 *are given by the quantities in the assumption.*

Proof sketch.

The second inequality follows from the smoothness of the objective. The first inequality follows from the smoothness of $d\hat{\theta}(\vec{w})/d\vec{w}$. □

Corollary

Under standard conditions, Assumption 1 holds for fixed constants with probability approaching one for $N \to \infty$. *Then Theorem 1 applies with probability approaching one as* $N \to \infty$.

Microcredit.

Study case	Original estimate	Target change	Refit estimate	Observations dropped
Bosnia	37.534 (19.780)	Sign change	**-2.226 (15.628)**	14 = 1.17%
		Significance change	**43.732 (18.889)***	1 = 0.08%
		Significant sign change	**-34.929 (14.323)***	40 = 3.35%
Ethiopia	7.289 (7.893)	Sign change	**-0.053 (2.513)**	1 = 0.03%
		Significance change	**15.356 (7.763)***	45 = 1.45%
		Significant sign change	**-8.755 (1.852)***	66 = 2.12%
India	16.722 (11.830)	Sign change	**-0.501 (8.221)**	6 = 0.09%
		Significance change	**22.895 (10.267)***	1 = 0.01%
		Significant sign change	**-16.638 (7.537)***	32 = 0.47%
Mexico	-4.549 (5.879)	Sign change	**0.398 (3.194)**	1 = 0.01%
		Significance change	**-10.962 (5.565)***	14 = 0.08%
		Significant sign change	**7.030 (2.549)***	15 = 0.09%
Mongolia	-0.341 (0.223)	Sign change	**0.021 (0.184)**	16 = 1.66%
		Significance change	**-0.436 (0.220)***	2 = 0.21%
		Significant sign change	**0.361 (0.147)***	38 = 3.95%
Morocco	17.544 (11.401)	Sign change	**-0.569 (9.920)**	11 = 0.20%
		Significance change	**21.720 (11.003)***	2 = 0.04%
		Significant sign change	**-18.847 (9.007)***	30 = 0.55%
Philippines	66.564 (78.127)	Sign change	**-4.014 (57.204)**	9 = 0.81%
		Significance change	**138.929 (66.880)***	4 = 0.36%
		Significant sign change	**-122.494 (49.409)***	58 = 5.21%

Table: Microcredit regressions for the profit outcome. The "Refit estimate" column shows the result of re-fitting the model removing the Approximate Most Influential Set. Stars indicate significance at the 5% level. Refits that achieved the desired change are bolded.

Cash transfers.

Study case	Original estimate	Target change	Refit estimate	Observations dropped
Poor, period 10	33.861 (4.468)*	Sign change	-**2.559 (3.541)**	697 = 6.63%
		Significance change	**4.806 (3.684)**	435 = 4.14%
		Significant sign change	-**9.416 (3.296)***	986 = 9.37%
Non-poor, period 10	21.493 (9.405)*	Sign change	-**0.573 (6.750)**	30 = 0.70%
		Significance change	**16.262 (8.927)**	3 = 0.07%
		Significant sign change	-10.845 (6.467)	92 = 2.16%

Table: Cash transfers results for the final study period. The "Refit estimate" column shows the result of re-fitting the model removing the Approximate Most Influential Set. Stars indicate significance at the 5% level. Refits that achieved the desired change are bolded.

Conclusion: Related work and future directions

Influence function

The present work is based on the *empirical influence function*. Consider:

- True, unknown distribution function $F_\infty(x) = p(X \le x)$
- Empirical distribution function $\hat{F}(x) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}\left(x_n \le x\right)$
- A statistical functional $T(F)$.

Influence function

The present work is based on the *empirical influence function*. Consider:

- True, unknown distribution function $F_\infty(x) = p(X \le x)$
- Empirical distribution function $\hat{F}(x) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}(x_n \le x)$
- A statistical functional $T(F)$.

We estimate with $T(F_\infty)$ with $T(\hat{F})$.
Sample means are an example:

$$T(F) := \int x \, F(\mathrm{d}x).$$

Z-estimators are, too:

$$T(F) := \theta \text{ such that } \int G(\theta, x) F(\mathrm{d}x) = 0.$$

Influence function

The present work is based on the *empirical influence function*. Consider:

- True, unknown distribution function $F_\infty(x) = p(X \le x)$
- Empirical distribution function $\hat{F}(x) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}\left(x_n \le x\right)$
- A statistical functional $T(F)$.

Form an (infinite-dimensional) Taylor series expansion at some F_0:

$$T(F) = T(F_0) + T'(F_0)(F - F_0) + \mathrm{residual}.$$

When the derivative operator takes the form of an integral

$$T'(F_0)\Delta = \int \psi(x; F_0)\Delta(\mathrm{d}x)$$

then $\psi(x; F_0)$ is known as the *influence function*.
Where to form the expansion? There are at least two reasonable choices:

- The limiting influence function $\psi(x, F_\infty)$
- The empirical influence function $\psi(x, \hat{F})$

Influence function

- The limiting influence function (LIF) $\psi(x, F_\infty)$
  - Used in a lot of classical statistics [Mises, 1947, Huber, 1981, Hampel, 1986, Bickel et al., 1993]
  - Unobserved, asymptotic
  - Requires careful functional analysis [Reeds, 1976]
- The empirical influence function (EIF) $\psi(x, \hat{F})$
  - The basis of the present work (also [Giordano et al., 2019b,a])
  - Computable, finite-sample
  - Requires only finite-dimensional calculus

Typically the *semantics* of the EIF derive from study of the LIF.
Example: $\frac{1}{N}\sum_{n=1}^{N}(N\psi_n)^2 \approx \mathrm{Var}\left(\sqrt{N}\phi(\hat{\theta})\right)$.

But the EIF measures what happens when you perturb the data at hand.

Other data perturbations will admit an analysis similar to ours!

Local robustness

The present work is an application of *local robustness*. Consider:

- Model parameter λ (e.g., data weights $\lambda = \vec{w}$)
- Set of plausible models $\mathcal{S}_\lambda$ (e.g. $\mathcal{S}_\lambda = W_\alpha$)
- Estimator $\hat{\theta}(x, \lambda)$ for data x and $\lambda \in \mathcal{S}_\lambda$ (e.g. a Z-estimator)

Global robustness: $\left(\inf_{\lambda \in \mathcal{S}_\lambda} \hat{\theta}(x, \lambda), \sup_{\lambda \in \mathcal{S}_\lambda} \hat{\theta}(x, \lambda)\right)$ (Hard in general!)

Local robustness: $\left(\inf_{\lambda \in \mathcal{S}_\lambda} \hat{\theta}^{lin}(x, \lambda), \sup_{\lambda \in \mathcal{S}_\lambda} \hat{\theta}^{lin}(x, \lambda)\right)$

...where $\hat{\theta}^{lin}(x, \lambda) := \hat{\theta}^{lin}(x, \lambda_0) + \left.\frac{\partial \hat{\theta}^{lin}(x, \lambda)}{\partial \lambda}\right|_{\lambda_0} (\lambda - \lambda_0)$.

Many variants are possible!

- Cross-validation [Giordano et al., 2019b]
- Prior sensitivity in Bayesian nonparametrics [Giordano et al., 2021]
- Model sensitivity of MCMC output [Giordano et al., 2018]
- Frequentist variances of MCMC posteriors (in progress)

Conclusion

- You may be concerned if you could reverse your conclusion by removing a $\lfloor \alpha N \rfloor$ datapoints, for some small α.

Conclusion

- You may be concerned if you could reverse your conclusion by removing a $\lfloor \alpha N \rfloor$ datapoints, for some small α.
- Robustness to removing a $\lfloor \alpha N \rfloor$ datapoints is principally determined by the signal to noise ratio, does not disappear asymptotically, and is distinct from (and typically larger than) standard errors.

Conclusion

- You may be concerned if you could reverse your conclusion by removing a $\lfloor \alpha N \rfloor$ datapoints, for some small α.
- Robustness to removing a $\lfloor \alpha N \rfloor$ datapoints is principally determined by the signal to noise ratio, does not disappear asymptotically, and is distinct from (and typically larger than) standard errors.
- Robustness to removing a $\lfloor \alpha N \rfloor$ datapoints is easy to check! We can quickly and automatically find an approximate influential set which is accurate for small α.

Conclusion

- You may be concerned if you could reverse your conclusion by removing a $\lfloor \alpha N \rfloor$ datapoints, for some small α.
- Robustness to removing a $\lfloor \alpha N \rfloor$ datapoints is principally determined by the signal to noise ratio, does not disappear asymptotically, and is distinct from (and typically larger than) standard errors.
- Robustness to removing a $\lfloor \alpha N \rfloor$ datapoints is easy to check! We can quickly and automatically find an approximate influential set which is accurate for small α.
- In the present work, we studied data dropping. But we provide a framework for studying many other robustness questions, both to data and model perturbations.

Links and references

Tamara Broderick, Ryan Giordano, Rachael Meager (alphabetical authors)
"An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?"
https://arxiv.org/abs/2011.14999

M. Angelucci, D. Karlan, and J. Zinman. Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco. *American Economic Journal: Applied Economics*, 7(1):151–82, 2015.

P. Bickel, C. Klaassen, Y. Ritov, and J Wellner. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

R. Giordano, T. Broderick, and M. I. Jordan. Covariances, robustness and variational Bayes. *The Journal of Machine Learning Research*, 19(1):1981–2029, 2018.

R. Giordano, M. I. Jordan, and T. Broderick. A higher-order Swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*, 2019a.

R. Giordano, W. Stephenson, R. Liu, M. I. Jordan, and T. Broderick. A swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR, 2019b.

R. Giordano, R. Liu, M. I. Jordan, and T. Broderick. Evaluating sensitivity to the stick-breaking prior in Bayesian nonparametrics. 2021.

F. Hampel. *Robust statistics: The approach based on influence functions*, volume 196. Wiley-Interscience, 1986.

P. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Weather modification*, volume 5, page 221. Univ of California Press, 1967.

P. Huber. *Robust Statistics*. John Wiley & Sons, New York, 1981.

R. Mises. On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3):309–348, 1947.

J. Reeds. *On the definition of von Mises functionals*. PhD thesis, Statistics, Harvard University, 1976.

L. Stefanski and D. Boos. The calculus of M-estimation. *The American Statistician*, 56(1):29–38, 2002.